Sentimental Analysis of Hotel Reviews from TripAdvisor

Vaibhav Singh[1], Aayushi Mahajan[2], Deepanshi Chaudhary[3]

[1]*Student, Computer Science and Engineering Department, ABES Engineering College, Ghaziabad, India*
[2]*Student, Computer Science and Engineering Department, ABES Engineering College, Ghaziabad, India*
[3]Asst. Professor, *Computer Science and Engineering Department, ABES Engineering College, Ghaziabad, India*

---***---

Abstract - *Sentiment Analysis as the name suggests is a machine learning technique that allows machines to read through human emotions. Allowing machines to read and understand through human emotions and extract useful insights through them is a vital resource for many businesses to grow and develop in their field. Hotel reviews collected from the guests can be classified into three subclasses i.e. positive, negative, or neutral and therefore we can analyze the sentiment of the customer. A crisp and to the point analysis of these reviews are vital for the quality control of the hotel services itself. For the classification for the hotel review data crawled from TripAdvisor, it is being observed that the statistical learning algorithms can provide better results than the human-generated sentiment analysis standard. For our classification, we will explore Naïve Bayes Classifier (NB), a machine learning library called Text Blob to calculate the subjectivity and polarity. To extract the frequency of words from the reviews we have used the Term Frequency -Inverse Document Frequency (TFIDF) approach. At last, we will conclude our approach by giving the accuracy of the model and discuss future scope.*

***Key Words*: Sentiment Analysis; Text Classification, Natural Language Processing; Aspect Extraction; Opinion Mining.**

1. INTRODUCTION

The importance of online reviews plays a vital role: **it is the key to your hotel standing on the online portals -** a precious license for delighted guests, which in turn leads to greater business and increased revenue outcomes. Precise management for your brand on online portals will reassure potential customers and **motivate them to opt for the hotel without a second thought in their mind.** Getting the reviews classified to gain insights from it, is now an **important part of the hotel business**. Reviews tell the story of how the customer feels about the services which the hotel is providing. The positive reviews can also be used to promote the good efforts of the hotel **just as important** as to take the negative reviews into account.

Sentiment analysis helps to improve the hotel business in several ways, from preventing a shrinking reputation in the market to understanding how the guests feel about their facility. Since there are tons of reviews available through different online platforms analyzing by themselves is no longer accountable for hotel businesses. They require accurate, reliable, fast, and efficient automated systems that can provide better findings to empower business decisions.

Sentiment analysis is indeed required to automate the process of determining whether a review expresses a positive, negative, or neutral opinion about the hotel and its services. With the help of sentiment analysis, hotels can save limitless time labeling customer data such as reviews, ratings, and comments on social media platforms. Sentiment analysis is required by the hotels to monitor their brand value on online portals, and gain information from customer feedback, and in turn, apply them to improve themselves.

For calculating the polarity score, many rule-based methods are defined for sentiment analysis using Natural Language Processing (NLP) techniques such as parsing, stemming, and tokenization alongside manually constructed rules. Firstly, it is necessary to define two lists of differing word parameters (For example positive words such as decent, greatest, lovely and negative words such as worse, horrible, poor, etc.). After this a rule-based system can be feed to the lists of predefined words, the system will give the count the of positive, negative and neutral sentiments that appears in the review, and will return a negative sentiment if it finds more negative words than positive words, and vice versa.

2. RELATED WORK

After referring to various research papers, it was evident that Sentiment Analysis is a fruitful task. It can be used in many applications in various fields such as for enhancing the customer's experience, reviving brand value, monitoring comments from social media, etc. Further, numerous features of Sentiment analysis are extracted for the attainment of deep insights into what's happening across your business channels [4].

During this work, we have observed that there is a massive growth in the numbers that are focusing on topics relevant to opinion mining, text classification, and sentiment analysis in recent years. As per the data available on the internet, there are nearly 11 thousand papers on relevant topics that have been published, and the point to acknowledge is that more than 95% of them appeared after the year 2004 which indeed makes sentiment analysis one of fastest-growing area for the researchers. In the present scenario, most of the papers focus on the research articles of sentiment analysis, which states that the topic is getting attention in the general public as well. Figure 1 shows the evolution of the research field related to systematic review.

We have also got an interesting observation that the university studies which measure public opinions were mostly taken place during and after World War-II and the motivation is highly legislative. This outburst of current interest on the topic happened only in the early-mid of 2000s, and it is mostly focused on the user sentiments available through the review and comments on the online portals, Since that time, the role of sentiment analysis has been prolonged to numerous other areas such as the prediction of Brand Value of products and amazingly reactions to terrorist attacks as well. Also, much corresponding research ongoing in the field of sentimental analysis and natural language processing has solved a lot of problems which adds to the use of sentiment analysis such as sarcasm detection and multilingual support. To add up more in this regard of getting human emotions analyzed through different approaches are were machine advancements are gradually doing a lot better than a simple human approach to the more complex gradation of emotions for getting them labeled as positive, negative and neutral emotions such as anger and pain.

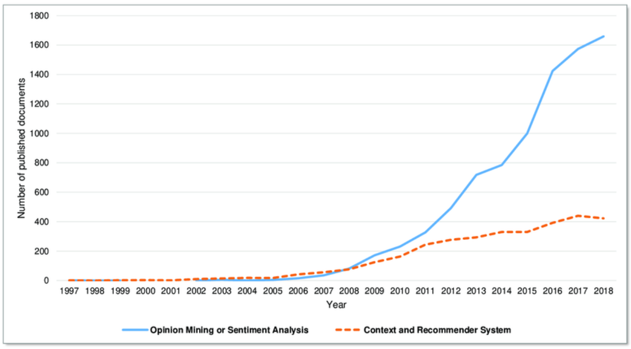


Figure 1. The evolution of the research field related to systematic review.

3. OVERVIEW OF THE SYSTEM

This section talks about the various technologies that are to be taken into account while running through the system.

A. Libraries Used

Python has a vast reserve of inbuilt standard libraries which includes areas like web services tools, string operation, data analysis, and machine learning, etc. The complex programming tasks can be dealt with ease using these inbuilt libraries as it reduces the size of code with many inbuilt functions that do the job pretty well for its user.

- *Scikit-learn*: - Scikit-learn is one of the most useful libraries that python offers. It has various statistical learning algorithms such as regression models (linear regression, logistic regression), SVM's, random forest for classification tasks and k-means for clustering, etc.

- *NumPy:* - The NumPy library in python is used for scientific computing and array manipulation. It can perform different operations such as indexing of an array, sequencing, and slicing, etc.

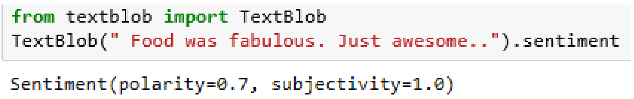
- *Pandas*: - The Pandas library in python is used for structuring, manipulating, and organizing data in a tabular structure called the data frame which is further used for data analysis.
- *TextBlob*: - The TextBlob library in python is used to process text data. It supports simple API for natural language processing (NLP) which does the job of calculating the polarity and subjectivity score pretty well.
- *NLTK*:-The NLTK library in python stands for Natural Language Toolkit which is used to pre-process human language so that machines can perform statistical analysis more accurately which is known as natural language processing (NLP).

B. Techniques

Sentiment analysis techniques can be classified into three major categories such as 1. Statistical methods, 2. Knowledge-based methods, and 3. Hybrid techniques. Statistical methods are also known as evolutionary approaches the main concept behind the approach is to find the mutual relationship between two words sharing the same context. It leverages some sort of mathematical representation of the text corpus. A simple approach to statistical methods is to sort say 500 words from a list that occurred most frequently in positive texts only excluding the negative ones and vice versa. Then, we can train a model and check whether there are more positive or negative words. This approach will be statistical since we are not leveraging linguistic insight (generally the distinction from a linguistic or lexical model). The goal of Knowledge-based methods is to extract knowledge by classifying text based on categories explicitly present in words such as awesome, sad, happy, unfortunate, and poor, etc. The knowledge bases also extract knowledge from unobvious words such as 'sympathy' assigned with particular emotions. The hybrid approach as the name suggests it constitutes both the above methods i.e. the statistical learning approach and the knowledge-based method for calculating the polarity scores. The reason for combining is to gain high accuracy and stability of the system at the same time.

- TextBlob Sentiment: Calculating Polarity and Subjectivity.

  The TextBlob library from python is widely used for sentiment analysis and is built on the top of NLTK. The text blob sentiment calculates the sentiment polarity and subjectivity as shown in Figure 2.



```
from textblob import TextBlob
TextBlob(" Food was fabulous. Just awesome..").sentiment

Sentiment(polarity=0.7, subjectivity=1.0)
```

Figure 2. Calculating the sentiment polarity and subjectivity.

This tells us that the English phrase “Food was fabulous. Just awesome.” has a polarity of about 0.7 states that it is positive sentiment and subjectivity of about 1.0, stating it is highly subjective.

- Sentiment Analysis Intuition behind Subjectivity and Polarity:

  The sentiment of a phrase will return a tuple in the form of Sentiment (polarity, subjectivity). A polarity score is a floating-point number that lies in the range [-1, 1] where values tending more towards 1 means positive statement 0 means neutral, and value tending towards -1 means negative sentiment.

  Subjectivity is also a floating-point number lies in the range [0, 1] where values tending towards 0 means objective statement and values tending towards 1 means subjective statement. Subjectivity refers to someone's personal feelings, opinions whereas Objective refers to factual statements.

- Text Classification (Naïve Bayes):

  The Naive Bayes Classifier is widely used for semantic sentiment analysis. It is a non-trivial but syntactic approach i.e. morphological or word-level analysis can be performed. The classifier will take two separate classes of positive words and negative words then it constructs a conditional density table based on the frequencies of positive words and negative words. Predicting an arbitrary statement represented by a bag of words is formed by log-summing the probabilities in class conditional tables. It will return two such summations, one for the positive class and the other for the negative class and compute sum score to make predictions.

  Bayes’ Rule Applied to Documents and Classes:

  For a document d and a class c,

$$P(c \mid d) = \frac{P(d \mid c)P(c)}{P(d)}$$

Naïve Bayes Intuition I:

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}}\, P(c \mid d)$$

MAP is "maximum a posteriori" = most likely class

$$= \underset{c \in C}{\operatorname{argmax}} \frac{P(d \mid c)P(c)}{P(d)}$$

Bayes Rule

$$= \underset{c \in C}{\operatorname{argmax}}\, P(d \mid c)P(c)$$

Dropping the denominator

Naïve Bayes Intuition II:

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}}\, P(d \mid c)P(c)$$

$$= \underset{c \in C}{\operatorname{argmax}}\, P(x_1, x_2, \ldots, x_n \mid c)P(c)$$

Document d represented as features x1..xn

Naïve Bayes Intuition III:

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}}\, P(x_1, x_2, \ldots, x_n \mid c)P(c)$$

$O(|X|^n \bullet |C|)$ parameters

How often does this class occur?

Could only be estimated if a very, very large number of training examples was available.

We can just count the relative frequencies in a corpus

C. Dataset

The Dataset is crawled from TripAdvisor.com using a method called web scraping. Web Scrapping is a method used for extracting a large amount of data from the web. We extracted hotel review data for Crowne Plaza Hotel Mayur Vihar, New Delhi, and stored it in .csv format with the Beautiful Soup a web scrapping library provided by python. The dataset contains two columns i.e. review column and dates to reviews with a total of 738 rows.

4. PROPOSED METHODOLOGY

A. General Method

In our work, we will perform sentiment analysis in Python, because it performs the task pretty well and provides us with crisp and beautiful visualizations in the form of graphs and various plots.

There are 5 steps to perform sentiment analysis on a given set of data as shown in Figure 3.

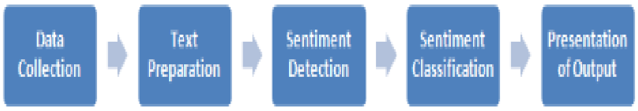


Figure 3. A flow diagram representation of Sentiment Analysis.

B. Execution Process

C. *Data Collection*- The dataset is extracted from TripAdvisor.com for Crowne Plaza Hotel Mayur Vihar, New Delhi, and stored into .csv format.

D. *Data Preprocessing*- The data preprocessing is nothing but filtering and clean the data in the following steps such that data is ready for further processing. Data preprocessing is done to obtain cleaner data which in turn will provide ease of processing the data further to obtain meaningful results, steps preprocess the data is shown in Figure 4.

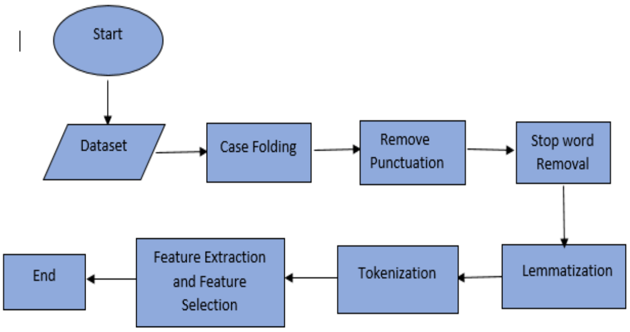
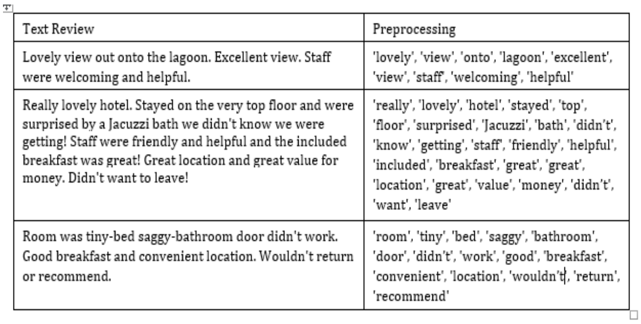


Figure 4. A flow diagram representation of Data Preprocessing

Illustration of preprocessing of the Text Data:



Text Review	Preprocessing
Lovely view out onto the lagoon. Excellent view. Staff were welcoming and helpful.	'lovely', 'view', 'onto', 'lagoon', 'excellent', 'view', 'staff', 'welcoming', 'helpful'
Really lovely hotel. Stayed on the very top floor and were surprised by a Jacuzzi bath we didn't know we were getting! Staff were friendly and helpful and the included breakfast was great! Great location and great value for money. Didn't want to leave!	'really', 'lovely', 'hotel', 'stayed', 'top', 'floor', 'surprised', 'Jacuzzi', 'bath', 'didn't', 'know', 'getting', 'staff', 'friendly', 'helpful', 'included', 'breakfast', 'great', 'great', 'location', 'great', 'value', 'money', 'didn't', 'want', 'leave'
Room was tiny-bed saggy-bathroom door didn't work. Good breakfast and convenient location. Wouldn't return or recommend.	'room', 'tiny', 'bed', 'saggy', 'bathroom', 'door', 'didn't', 'work', 'good', 'breakfast', 'convenient', 'location', 'wouldn't', 'return', 'recommend'

3. Sentiment Detection- In this Phase, the reviews and opinions are further analyzed to calculate the sentiment polarity and subjectivity. In our case, it is done by using the Python inbuilt library TextBlob.

4. Sentence Classification- Once the subjectivity and polarity scores are being generated the sentiments can be classified as positive, negative, and neutral.

5. Output Presentation- The reason output presentation is an essential task is to provide meaning to our results. The information gained from the analysis is required to be presented in such a manner that it reflects meaningful information from the unstructured data by visualizing and plotting the results.

5. Evaluation and Results

A. Evaluation

We evaluated the analysis system on a corpus of 738 hotel reviews crawled from the web. For the evaluation, these segments were manually classified concerning their polarity, including the neutral polarity besides positive and negative ones.

B. Results

The results were as out of 738 reviews we found that 97.3 percent of the total reviews were listed as positive, 2.6 percent of the total reviews were listed as negative and 0.1 percent of the reviews are neutral which are further demonstrated with the help of bar to provide a better insight as shown in Figure 5.

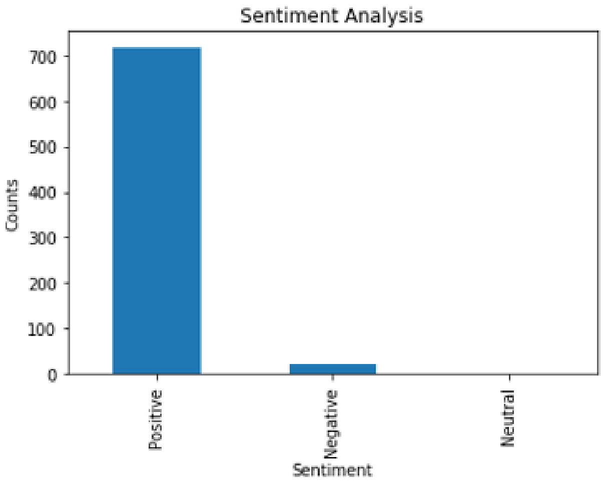


Figure 5. Reviews classified into three categories Positive, negative, and neutral.

To further demonstrate and visualize the result a word cloud is formed from the reviews as shown the Figure 6. A word cloud gives the frequency of words used in a particular sentence in our case reviews and describes its importance in the form of an image.

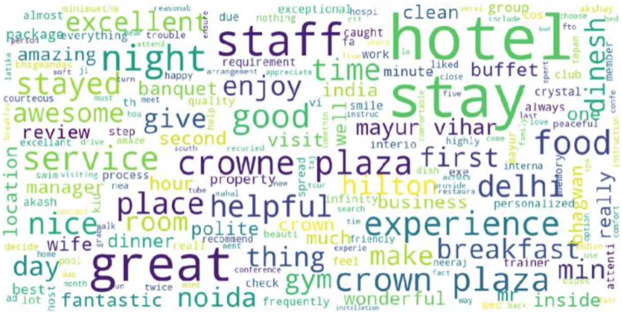
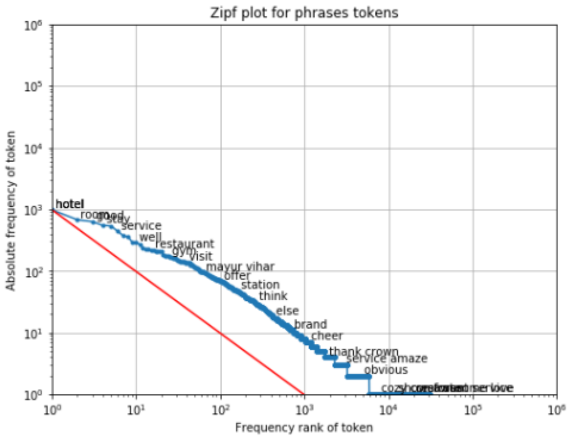
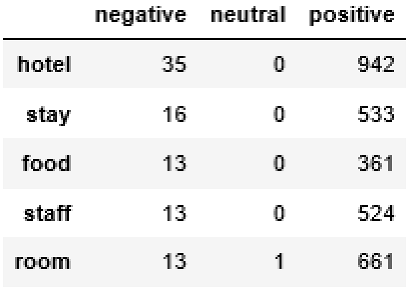


Figure 6. A Word Cloud demonstrating the frequency and importance of words appeared in the reviews.

The words carrying higher frequency are analyzed alone such as hotel, staff, food, etc. to gain meaningful insight into these aspects by counting their frequencies in reviews and labeling them positive, negative and neutral based on the above results. So that we can know which aspect of the hotel most and least important in our overall analysis which is shown in Figure 7.





	negative	neutral	positive
hotel	35	0	942
stay	16	0	533
food	13	0	361
staff	13	0	524
room	13	1	661

Figure 7. Analysis based upon the aspects.

6. CONCLUSIONS

Sentiment analysis for the hotel reviews has been carried out labeling reviews as positive sentiments which include a word like- happy, amazing, tasty, nice, pretty as well as negative sentiments which include words bad, disgusting, sad, and disappointed, etc. The whole point of the analysis is to provide suitable recommendations to the customers to select the best available option and to the business owner for successful decision making, using sentiment-based results, and implying sentiments. Moreover, the sentiment analysis in this work has been applied to determine the attitude of customers through online feedbacks given by them on hotel services, food, staff, and ambiance of the respective hotel.

REFERENCES

[1] Multi-language sentimental analysis for hotel reviews Maleerat Sodanila MATEC Web ICMIE 2016

[2] Sentiment Analysis on Chinese Hotel Reviews with Doc2Vec and Classifiers 2018 IEEE 3rd Advanced Information Technology, Electronic, and Automation conference.

[3] Sentiment Analysis of Review Datasets using Naïve Bayes' and K-NN Classifier.

[4] Automated Aspect Extraction and Aspect-Oriented Sentiment Analysis on Hotel Review Datasets.

[5] Sentiment analysis on hotel reviews using Multinomial Naïve Bayes classifier. IOP Conf. Series: Journal of Physics: Conf. Series 1192 (2019) 012024.

[6]Assessing the helpfulness of online hotel reviews: A classificationbased approach Pei-Ju Lee, Ya-Han Hu□, Kuan-Ting Lu.

[7]Understanding Online Hotel Reviews through Automated Text Analysis: Shawn Mankad, Hyunjeong "Spring" Han, Joel Goh, Srinagesh Gavirneni.

[8]Opinion mining from online hotel review: Ya-Han Hu a, Yen-Liang Chen b, Hui-Ling Chou.